

# Variable Selection for Latent Dirichlet Allocation

Dongwoo Kim, Yeonseung Chung, and Alice Oh  
KAIST, Korea

March 3, 2013

## Abstract

In latent Dirichlet allocation (LDA), topics are multinomial distributions over the entire vocabulary. However, the vocabulary usually contains many words that are not relevant in forming the topics. We adopt a variable selection method widely used in statistical modeling as a dimension reduction tool and combine it with LDA. In this variable selection model for LDA (vsLDA), topics are multinomial distributions over a subset of the vocabulary, and by excluding words that are not informative for finding the latent topic structure of the corpus, vsLDA finds topics that are more robust and discriminative. We compare three models, vsLDA, LDA with symmetric priors, and LDA with asymmetric priors, on heldout likelihood, MCMC chain consistency, and document classification. The performance of vsLDA is better than symmetric LDA for likelihood and classification, better than asymmetric LDA for consistency and classification, and about the same in the other comparisons.

## 1 Introduction

Latent Dirichlet allocation (LDA) [2], a widely used topic model, decomposes a corpus into a finite set of topics. Each topic is a multinomial distribution over the entire vocabulary, which is typically defined to be the set of all unique words with an optional step of removing stopwords and high frequency words. Even with the preprocessing step, the vocabulary will almost certainly contain words that do not contribute to the underlying topical structure of the corpus, and those words may interfere with the model’s ability to find topics with predictive and discriminative power. More importantly, one cannot be sure whether and how much the vocabulary influences the topics inferred, and there is not a systematic way to compare different vocabularies for a given corpus. We relax the constraint that the vocabulary must be fixed a priori and let the topic model consider any subset of the vocabulary for representing the topics.

We propose a model-based variable selection [5, 9] for LDA (vsLDA) that combines the process of identifying a relevant subset of the vocabulary with

the process of finding the topics. Variable selection has not been studied in depth for LDA or any other topic model, but three models, HMM-LDA [7], sparseTM [22], and SWB [4] achieve a similar effect of representing the topics with a subset of the vocabulary. HMM-LDA [7] models the short- and long-range dependencies of the words and thus identifies whether words are generated from the syntactic (non-topic) or the semantic (topic) class. SparseTM [22] aims to decouple sparsity and smoothness of the word-topic distribution and thereby excludes some words from each topic. SWB separates *word tokens* into the general and specific aspects, and it is probably the most similar work to ours in that it also globally excludes words from forming the topics. However, SWB excludes *word tokens*, whereas vsLDA excludes *word types*. By looking at the word types, we can replace the necessary but arbitrary step of deciding the vocabulary for forming the topics, which usually includes the removal process of useless words. Such process typically uses a list of stop words and corpus-dependent infrequent and highly frequent words, and in this work, we show the inadequacy of such preprocessing approach to variable selection. We can also view this problem of variable selection as a type of model selection along the vocabulary dimension. Model selection has been well studied for the topic dimension with nonparametric topic models [19, 22] but not for the vocabulary dimension.

This paper is organized as follows. We first describe our vsLDA model for selecting informative words. We derive an approximate algorithm for posterior inference on the latent variables of interest in vsLDA based on Markov Chain Monte Carlo and Monte Carlo integration. We demonstrate our approach on a synthetic dataset to verify the correctness of our model. Then we run our model on three real-world datasets and compare the performance with LDA with symmetric priors (symLDA) and LDA with asymmetric priors (asymLDA). We show that vsLDA finds topics with better predictive power than symLDA and more robustness than asymLDA. We also find that vsLDA reduces each document into more discriminating subdimensions and hence outperforms the other models for document classification.

## 2 Variable Selection for LDA (vsLDA)

LDA is typically used with a preprocessing step of removing stopwords and the words that occur frequently throughout the corpus. The rationale is that the words pervading the corpus do not contribute to but hinder the process of discovering a latent topic structure. This frequency-based preprocessing step excludes the words a priori independent of constructing the latent topics. However, we cannot be certain whether the excluded words are truly non-informative for topic construction. Also, the same uncertainty applies to the included words. Here, we propose a new LDA model where the word selection is conducted simultaneously while discovering the latent topics. The proposed approach combines a stochastic search variable selection [5] with LDA, providing an automatic word selection procedure for topic models.

Suppose we have a vocabulary with size  $V$  with or without any preprocessing. In a typical topic model, topics are defined on the entire vocabulary and assumed to be Dirichlet-distributed on  $V - 1$  simplex, i.e.,

$$\phi_k \sim \text{Dir}(\beta \mathbf{1}), \quad k \in \{1, 2, 3, \dots, K\}, \quad (1)$$

where  $\mathbf{1}$  is  $V$ -dimensional vector of 1s and  $K$  is the number of topics. Our assumption is that the vocabulary is divided into two mutually exclusive word sets; one includes informative words for constructing topics, and the other contains non-informative words. Also, the topics are assumed to be defined only on the informative word set and distributed as

$$\phi_k \sim \text{Dir}(\beta \mathbf{s}), \quad k \in \{1, 2, 3, \dots, K\}, \quad (2)$$

where  $\mathbf{s} = (s_1, \dots, s_V)$  and  $s_j$  is an indicator variable defined as

$$s_j = \begin{cases} 1, & \text{word } j \text{ is a informative word,} \\ 0, & \text{word } j \text{ is a non-informative word.} \end{cases} \quad (3)$$

In other words,  $\mathbf{s}$  specifies a smaller simplex with a dimension  $\sum_{j=1}^V s_j - 1$  for the informative word set. Not knowing a priori whether a word is informative or non-informative, we assume  $s_j \sim \text{Bernoulli}(\lambda)$  to incorporate uncertainty in informativity of words.

Now, we describe the generative process for vsLDA which includes the steps for dividing the entire vocabulary into an informative word set and a non-informative word set (step 1) and determining the membership of a word token either as one of the topics or as the non-informative word set (step 4(b)).

1. For each word  $j \in \{1, 2, \dots, V\}$ , draw word selection variable  $s_j \sim \text{Bernoulli}(\lambda)$
2. For each topic  $k \in \{1, 2, \dots, K\}$ , draw topic distribution  $\phi_k \sim \text{Dir}(\beta \mathbf{s})$
3. For a non-informative words set, draw words distribution  $\psi \sim \text{Dir}(\gamma \mathbf{s}^c)$
4. For each document  $d \in \{1, 2, \dots, D\}$ :
  - (a) Draw topic proportion  $\theta_d \sim \text{Dir}(\alpha)$
  - (b) For  $i$ th word token, draw  $b_{di} \sim \text{Bernoulli}(\tau)$ :
    - i. If  $b_{di} = 1$ :
      - A. Draw topic assignment  $z_{di} \sim \text{Mult}(\theta_d)$
      - B. Draw word token  $w_{di} \sim \text{Mult}(\phi_{z_{di}})$
    - ii. else
      - A. Draw word token  $w_{di} \sim \text{Mult}(\psi)$

From the generative process for a corpus, the likelihood of the corpus is

$$\begin{aligned}
& P(W, \Theta, \Phi, \psi, \mathbf{z}, \mathbf{b}, \mathbf{s} | \alpha, \beta, \gamma, \lambda, \tau) \\
&= \prod_{d=1}^D p(\theta_d | \alpha) \prod_{d=1}^D \prod_{\{i: b_{di}=1\}} \{p(w_{di} | \phi_{z_{di}}) p(z_{di} | \theta_d)\} \\
&\times \prod_{d=1}^D \prod_{\{i: b_{di}=0\}} p(w_{di} | \psi) \prod_{d=1}^D \prod_{i=1}^{N_d} p(b_{di} | \tau) \\
&\times \prod_{k=1}^K p(\phi_k | \beta, \mathbf{s}) p(\psi | \gamma, \mathbf{s}) \prod_{j=1}^V p(s_j | \lambda),
\end{aligned}$$

where  $N_d$  is the number of word tokens in document  $d$ .

Placing Dirichlet-multinomial conjugate priors over  $\Theta, \Phi, \psi$  naturally leads to marginalizing out these variables.

$$\begin{aligned}
& p(W, \mathbf{z}, \mathbf{b}, \mathbf{s} | \alpha, \beta, \gamma, \lambda, \tau) = \\
& \int_{\Theta} \prod_{d=1}^D \prod_{\{i: b_{di}=1\}} p(z_{di} | \theta_d) p(\theta_d | \alpha) d\Theta \\
& \times \int_{\Phi} \prod_{k=1}^K p(\phi_k | \beta, \mathbf{s}) \prod_{d=1}^D \prod_{\{i: b_{di}=1\}} p(w_{di} | \phi_{z_{di}}) d\Phi \\
& \times \int_{\psi} p(\psi | \gamma, \mathbf{s}) \prod_{d=1}^D \prod_{\{i: b_{di}=0\}} p(w_{di} | \psi) d\psi \\
& \times \prod_{j=1}^V p(s_j | \lambda) \prod_{d=1}^D \prod_{i=1}^{N_d} p(b_{di} | \tau) \tag{4} \\
&= \prod_{d=1}^D \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\prod_{k=1}^K \Gamma(n_{d\cdot}^k + \alpha_k)}{\Gamma(\sum_{k=1}^K n_{d\cdot}^k + \alpha_k)} \\
&\times \prod_{k=1}^K \frac{\Gamma(\sum_{\{j: s_j=1\}} \beta_j)}{\prod_{\{j: s_j=1\}} \Gamma(\beta_j)} \frac{\prod_{\{j: s_j=1\}} \Gamma(n_{d\cdot}^k + \beta_j)}{\Gamma(\sum_{\{j: s_j=1\}} n_{d\cdot}^k + \beta_j)} \\
&\times \frac{\Gamma(\sum_{\{j: s_j=0\}} \gamma_j)}{\prod_{\{j: s_j=0\}} \Gamma(\gamma_j)} \frac{\prod_{\{j: s_j=0\}} \Gamma(m_{d\cdot} + \gamma_j)}{\Gamma(\sum_{\{j: s_j=0\}} m_{d\cdot} + \gamma_j)} \\
&\times \lambda^{|\mathbf{s}|} (1 - \lambda)^{|V| - |\mathbf{s}|} \cdot \tau^{n_{\cdot\cdot}} (1 - \tau)^{m_{\cdot\cdot}}
\end{aligned}$$

where  $n_{dj}^k$  is a number of word tokens in the  $d$ th document with the  $j$ th word in the vocabulary assigned to the  $k$ th topic where  $s_j = 1$ , and  $m_{dj}$  is a number of word tokens in  $d$ th document with the  $j$ th word where  $s_j = 0$ . The dots represent the marginal counts, so  $m_{\cdot j}$  represents the number of word tokens of  $j$ th word across the corpus.

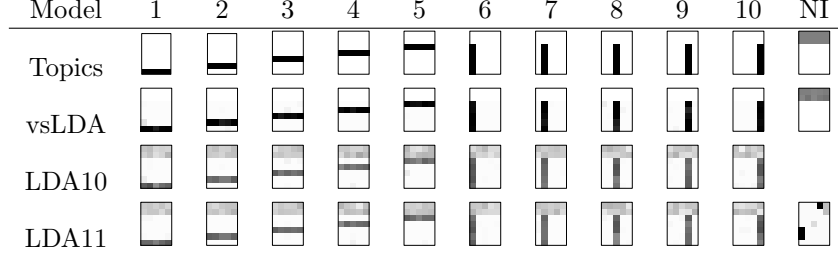


Figure 1: Result of the simulation study. The first row shows the topics used to generate the data. The second, third, and fourth rows show the inferred topics from vsLDA, LDA10, and LDA11, respectively. For all models, we use an asymmetric prior  $\alpha$  over the per-document topic proportions.

### 3 Approximate Posterior Inference

Deriving exact posterior distributions for the latent variables in vsLDA is intractable. We propose an MCMC algorithm to obtain posterior samples in order to make an approximate inference. Marginalizing over  $\Phi, \psi$ , and  $\Theta$ , the remaining latent variables in the joint likelihood are  $\mathbf{z}, \mathbf{b}$ , and  $\mathbf{s}$  in equation (4). Given the word selector  $\mathbf{s}$  and the observed data  $W$ ,  $\mathbf{b}$  is determined because  $b_{di} = 1$  for all informative word tokens and  $b_{di} = 0$  for all non-informative word tokens. Therefore, we sample only  $\mathbf{z}$  and  $\mathbf{s}$  through a collapsed Gibbs updating and a Metropolis updating relying on a Monte Carlo integration, respectively.

**Step 1 : Sampling  $\mathbf{z}$  :** Given  $W$  and  $\mathbf{s}$ , we sample  $z_{di}$  only for  $d$  and  $i$  such that  $w_{di} = j$  and  $s_j = 1$  (i.e., only for the word tokens taking the values in the informative word set). Letting  $\mathbf{z}_{-di} = \{z_{d'i'} : d' \neq d \text{ or } i' \neq i\}$ , the conditional distribution of  $z_{di}$  given  $\mathbf{z}_{-di}$ ,  $\mathbf{s}$ , and  $W$  is

$$p(z_{di} = k | W, \mathbf{z}, \mathbf{s}) \propto (n_{d\cdot}^k + \alpha) \frac{n_{\cdot w_{di}}^k + \beta}{n_{\cdot\cdot}^k + \beta \sum_{j=1}^V s_j} \quad (5)$$

which depends only on the number of informative words and the topic assignments of the other informative word tokens. This is a generalization of updating step for topic assignment in typical LDA models where the vocabulary size is fixed as  $V$  while it varies as  $\sum_{j=1}^V s_j$  in our model.

**Step 2 : Sampling  $\mathbf{s}$  :** We let  $\mathbf{z}^j = \{z_{di}; w_{di} = j\}$ ,  $\mathbf{z}^{-j} = \mathbf{z} \setminus \mathbf{z}^j$ , and  $\mathbf{s}^{-j} = \mathbf{s} \setminus s^j$ . We update  $s_j$  using a Metropolis step where  $s_j^{proposed}$  is accepted over  $s_j^{current}$  with a probability

$$\text{Min} \left\{ 1, \frac{\int p(W, \mathbf{z}^j, \mathbf{z}^{-j}, s_j^{proposed}, \mathbf{s}^{-j}) p^*(\mathbf{z}^j) d\mathbf{z}^j}{\int p(W, \mathbf{z}^j, \mathbf{z}^{-j}, s_j^{current}, \mathbf{s}^{-j}) p^*(\mathbf{z}^j) d\mathbf{z}^j} \right\} \quad (6)$$

where  $p^*(\mathbf{z}^j) = p(\mathbf{z}^{-j}, s_j, \mathbf{s}^{-j})$  is the conditional distribution of  $\mathbf{z}^j$  given all the others. If proposed or current  $s_j$  is 0,  $\mathbf{z}^j$  disappears in the joint likelihood as

$p(W, \mathbf{z}^{-j}, s_j = 0, \mathbf{s}^{-j})$  and we do not need to marginalize over  $\mathbf{z}^j$ . If proposed or current  $s_j$  is 1, marginalization over  $\mathbf{z}^j$  does not yield a closed form and we rely on a Monte Carlo integration as follows.

$$\begin{aligned} & \int p(W, \mathbf{z}^j, \mathbf{z}^{-j}, s_j = 1, \mathbf{s}^{-j}) p^*(\mathbf{z}^j) d\mathbf{z}^j \\ &= \frac{\sum_{u=1}^U p(W, \mathbf{z}^{j(u)}, \mathbf{z}^{-j}, s_j = 1, \mathbf{s}^{-j})}{U} \end{aligned} \quad (7)$$

where  $\mathbf{z}^{j(u)}$  is  $u$ th sample obtained from  $p^*(\mathbf{z}^j)$  as in equation (5). Once we obtain posterior samples of  $\mathbf{s}$ , inference about  $\mathbf{s}$  is done through

$$s_j = \operatorname{argmax}_{B < t \leq T} p(s_j^{(t)} | W, \text{Rest}^{(t)}) \quad (8)$$

where  $T$  is the total number of iterations and  $B$  is the burn-in count.

## 4 Simulation Study

We first verify the correctness of our model with a synthetic dataset. We generate the synthetic corpus as follows. We start with thirty-five words in the vocabulary, design ten topics such that each topic has five topic (informative) words with 0.2 probability each and zero probability for all other words. Then we add a non-informative set with ten words, that do not appear in any of the topics, with 0.1 probability each. The first row in Figure 1 shows these hand-crafted topics. As the figure shows, there are twenty-five informative words and ten non-informative words. Based on these topics, we generate 200 documents, 40 to 50 tokens each, with random topic proportions drawn from the Dirichlet distribution with a symmetric prior of 0.1. For each document, we set  $\tau$  to 0.6, which means 60% of word tokens are drawn from the topics, and 40% word tokens are drawn from the non-informative set.

With this synthetic corpus, we trained vsLDA with ten topics and LDA with ten (LDA10) and eleven (LDA11) topics. For the hyperparameters, we place an asymmetric  $\alpha$  prior over the document topic proportions, a symmetric  $\beta$  prior over the topic-word distributions, and a symmetric  $\gamma$  prior over the non-informative word distribution. These asymmetric  $\alpha$  and symmetric  $\beta$  priors can improve the performance of LDA compared to the widely used symmetric  $\alpha$  and  $\beta$  priors [20]. During the inference steps, we optimized these hyperparameters by using Minka’s fixed point iteration [17] except  $\gamma$  which we set to 1. Finally, we place Beta(1,1) priors over the hyperparameters  $\lambda$  and  $\tau$ .

Figure 1 shows the topics inferred by each model. The topics from LDA10 and LDA11 look less clear than the topics from vsLDA. By design, vsLDA infers the non-informative words and explicitly excludes them from the topics, so the resulting topics distribute all of the probability over the informative words, thereby discovering topics with clearer patterns. In this simulation, vsLDA

Dataset	# of docs	# of words	# of tokens	Stopwords
20NG	2,000	3,608	155,622	No
NIPS	1,740	2,613	104,069	Yes
SigGraph	783	2,808	54,804	No

Table 1: Dataset statistics. The *stopwords* column indicates whether stopwords were kept (yes) or removed (no).

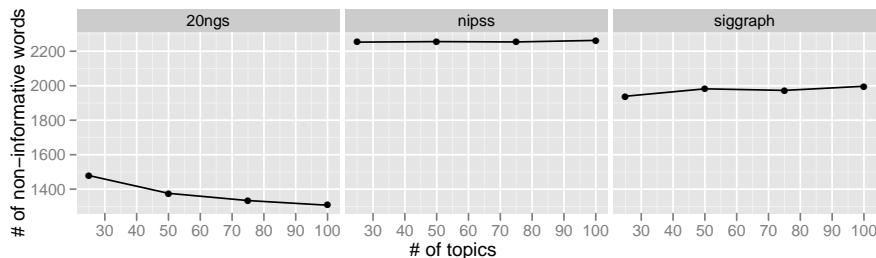


Figure 2: Change of the number of non-informative words over the number of topics ( $K$ ).

exactly captures the top five words in each topic, and correctly identifies, using equation (8), the set of non-informative words. LDA10 finds the top five words in each topic quite well. However, every topic identified by LDA10 distributes some probability over the non-informative words, so the topics are not clearly defined by the five topic words. One interesting point of discussion is that typically LDA with asymmetric  $\alpha$  priors are known to capture the common words of the corpus into a topic, so we expect that LDA11 would capture the non-informative words in its eleventh topic. However, topic number 11 in LDA11 actually captures an ambiguous distribution which has a sparse distribution over the words. However, if we adjust  $\tau$  to be smaller than 0.3, LDA11 captures the non-informative words into one topic as well.

## 5 Empirical Study

In this section, we analyze three corpora for comparing vsLDA with two variants of LDA using various evaluation metrics. The first two are abstracts collected from the proceedings of the ACM SigGraph conferences (SigGraph) and the proceedings of the NIPS conferences (NIPS), and the third dataset is from the five comp subcategories from the 20 newsgroup corpus (20NG). To show the performance of vsLDA for diverse settings, we test NIPS with stopwords kept and SigGraph and 20NG with stopwords removed. The detailed statistics of the three datasets are in Table 1. We compare three models: vsLDA, LDA with asymmetric priors for  $\theta$  (asymLDA), and LDA with symmetric priors for  $\theta$  (symLDA). Each model was run five times where each run includes 5,000

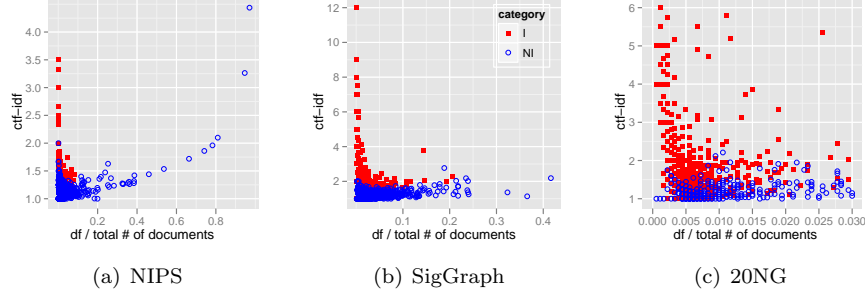


Figure 3: The scatter plot of  $ctf-idf$  versus relative  $df$  ( $rdf = df / total \# \text{ of documents}$ ) for the I words (red square) and the NI words (blue circle) inferred by vsLDA with  $K = 50$  for each corpus. Generally, I words tend to have higher  $ctf-idf$  and lower  $rdf$  than the NI words.

iterations with 3,000 burn-in samples and 100 iterations used as a thinning interval. Other parameters were optimized in the same way as the previous section.

**Characteristics of informative and non-informative words** We first describe the summary statistics of the informative (I) and the non-informative (NI) words found by vsLDA and explain the interesting patterns found. Figure 2 shows the pattern of how many words are found to be NI as we vary  $K$ , the number of topics. For NIPS and SigGraph, the number of NI words does not vary for  $K$  of 25, 50, 75, and 100. For 20NG, the number noticeably decreases as  $K$  increases. Further investigation is needed to explain this phenomenon, but as Table 3 shows, log likelihood of heldout data follows a similar trend, so one conjecture is that the optimal number of topics is related to the number of NI words for a given corpus. The proportion and the absolute number of the NI words clearly differ for each corpus. On average, vsLDA categorizes around 38%, 70%, and 86% of the words as NI for 20NG, SigGraph, and NIPS, respectively.

To compare the characteristics of the NI and the I words, we compute three summary statistics of the words: (1) corpus term frequency ( $freq$ ), (2) document frequency ( $df$ ), and (3) corpus  $tf-idf$  ( $ctf-idf$ ). Table 2 shows these statistics for ten words from 20NG and ten words from NIPS. To examine if any of the statistics are associated with word informativity, we ordered the words decreasingly by  $freq$  for 20NG and by  $ctf-idf$  for NIPS. Noting that there is no systematic pattern in the distribution of I and NI words in both orderings, we confirm that each statistic alone is not sufficient to distinguish the two classes of words inferred by vsLDA.

However, we found that the  $ctf-idf$  can be useful to quantify word informativity combined with the relative  $df$  ( $rdf = df / total \# \text{ of documents}$ ). Figure 3 shows a scatter plot of  $ctf-idf$  versus  $rdf$  for the I words (red square) and the NI words (blue circle) inferred from vsLDA. In particular, Figure 3(c) is a close-up



(a) 20NG				
word	freq	df	ctf-idf	category
subject	1,855	1,715	1.08	NI
re	970	915	1.06	NI
windows	918	356	2.58	I
writes	822	653	1.26	NI
file	766	206	3.72	I
article	686	537	1.28	NI
don't	597	394	1.52	NI
scsi	592	89	6.65	I
program	582	241	2.41	NI
drive	569	199	2.86	I

(b) NIPS				
word	freq	df	ctf-idf	category
the	6,764	1,524	4.44	NI
of	4,849	1,486	3.26	NI
speech	124	71	1.75	I
localization	19	11	1.73	I
is	1,791	1,042	1.72	NI
learning	647	397	1.63	NI
recurrent	77	58	1.33	I
hidden	132	106	1.25	I
feature	66	53	1.25	NI
can	493	396	1.24	NI

Table 2: Basic statistics for the I words and the NI words inferred by vsLDA with  $K = 50$  for 20NG and NIPS. The words are ordered decreasingly by *freq* (20NG) and by *ctf-idf* (NIPS), and neither ordering shows a systematic pattern of word informativity.

of the lower-left corner where most of the words are located for 20NG. As shown in Figure 3, the I words tend to show higher *ctf-idf* and lower *rdf* than the NI words, suggesting that the I words are the ones that appear in a few documents (low *rdf*) but with high frequency (high *ctf-idf*). For SigGraph with  $K = 50$ , the average *ctf-idf* of I words is 2.26 and the average *ctf-idf* of NI words is 1.27. The other corpora at all levels of  $K$  show the same pattern. As shown in Figure 3(c), many of the words show low *rdf*, and classification of these words mainly depends on the high/low *ctf-idf*.

In addition, Table 2 shows that the words normally categorized as stop words, such as “the” and “is” are correctly identified as NI, as are the words that do not distinguish topics, such as “learning” and “feature” in NIPS. We also found that the NIPS corpus contains 284 stopwords, and vsLDA identified 91% of them on average as NI words.

**Held-out likelihood** vsLDA divides the vocabulary into I and NI, two

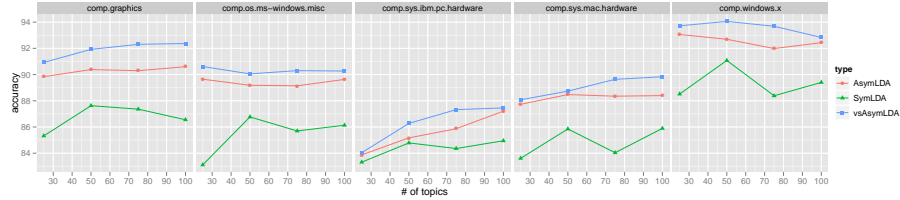


Figure 4: Average classification accuracies using 20NG. vsLDA outperforms symLDA and asymLDA on document classification.

mutually exclusive sets that are unpredictable given the basic word statistics. Now, we describe the performance of vsLDA using held-out likelihood which measures the model’s predictive performance for an unseen document based on the trained parameters. We split the corpus into a training set containing 90% of documents and a test set containing the rest. We compute held-out likelihoods using a left-to-right style sampler [21] with maximum a posteriori (MAP) estimators of parameters  $\hat{\phi}$ ,  $\hat{\psi}$ ,  $\hat{s}$ , and  $\hat{\tau}$ . The average word likelihoods are shown in Table 3, and these results are consistent with the values reported in other studies of LDA with symmetric priors [3] and LDA with asymmetric priors [20]. Overall, the held-out likelihoods of vsLDA are higher than symLDA and comparable to asymLDA. It is worth noting again that vsLDA excludes the NI words, which make up 40% to 80% of the vocabulary, from the topics, and it still performs comparable to asymLDA which uses all of the words for the topics. These results suggest that including the NI words in forming the topics does not contribute to the predictive power of the model. To further test vsLDA, we manually set the stop words as the non-informative words, trained the model (stopword-vsLDA), and computed held-out likelihoods. These heldout likelihoods were lower than vsLDA and asymLDA. These results verify that variable selection must be done within the model in combination with the topics, rather than as a preprocessing step.

**Classification** A topic model can be used for dimensionality reduction because it expresses each document as a finite mixture of topics. One way to verify the performance of a topic model is to perform classification tasks by using these reduced dimensions [2, 13]. We use the five subcategories of the 20NG dataset and classify the documents into the subcategories. We use the libSVM toolkit with linear kernels, performing one-vs-all classification on each category with ten-fold cross validation. Figure 4 shows the average accuracies of the classification results. Overall, vsLDA performs better than all others with a small difference between vsLDA and asymLDA. We can conclude that vsLDA reduces each document into more discriminating subdimensions by excluding the non-informative words.

**Similarity between multiple MCMC outputs with best matching algorithm** LDA with asymmetric priors tend to generate highly skewed distributions [20] where the model will capture several major topics well, but

(a) 20ng.comp				
	vsLDA	asymLDA	symLDA	
25	-7.04	-7.00	-7.35	
50	-6.94	-6.87	-7.37	
75	-6.88	-6.82	-7.35	
100	-6.84	-6.78	-7.36	

(b) SigGraph				
	vsLDA	asymLDA	symLDA	
25	-7.11	-7.04	-7.21	
50	-7.09	-7.02	-7.25	
75	-7.07	-7.00	-7.25	
100	-7.06	-6.99	-7.27	

(c) NIPS				
	vsLDA	asymLDA	symLDA	stopword-vsLDA*
25	-6.28	-6.25	-6.34	-6.32
50	-6.28	-6.25	-6.43	-6.34
75	-6.28	-6.25	-6.40	-6.34
100	-6.28	-6.25	-6.43	-6.35

Table 3:  $\log P(W^{\text{test}}|W)/N^{\text{test}}$  for various values of  $K$  for the three corpora. vsLDA performs comparable to asymLDA. For stopwords-vsLDA, we manually set the NI words with stopwords. stopwords-vsLDA performs comparable to symLDA but worse than vsLDA.

the other topics may be highly inconsistent over multiple MCMC outputs. In the experiments presented here, for instance, five major topics occupy more than 50% of word tokens in the corpus. This may pose a problem for cases where the inferred topics  $\hat{\phi}$ , not just the topic assignments for the word tokens, are important. Variation of information (VI) is one metric to evaluate the performance of clustering [16, 20], but VI is based on mutual information of the topic assignments of tokens, so the major topics of the asymmetric models will overtake the VI metric, thereby masking the inconsistencies of the minor topics.

In order to better measure the consistency of the model with respect to the topics  $\hat{\phi}$ , we propose a new similarity metric based on the best matching algorithm. First, based on the inferred  $K$  maximum a posterior (MAP)  $\hat{\phi}$ s for each MCMC output, we find the best matching  $K$  pairs that minimize the sum of symmetric KL-divergence with the Hungarian algorithm [6, 12]. If the model generates consistent topics over multiple runs, then the sum of the divergences will also be minimized. Table 4 shows the average divergences between the best matching pairs, and it shows that vsLDA finds more consistent topics than asymLDA and comparable results with symLDA. The inconsistencies of asymLDA can be attributed to the minor topics for which the corpus does not exhibit regular word-topic patterns. Although vsLDA may also generate skewed

(a) K=50			
	20ng.comp	NIPS	SigGraph
vsLDA	3.12	2.49	3.22
asymLDA	3.68	5.96	4.45
symLDA	2.74	2.21	2.68

(b) K=100			
	20ng.comp	NIPS	SigGraph
vsLDA	3.68	2.48	3.21
asymLDA	4.40	7.38	5.77
symLDA	3.00	2.20	2.69

Table 4: Average symmetric KL divergence between best matching topic pairs. vsLDA shows similar average divergences compared to symmetric LDA despite its asymmetric priors.

distributions, vsLDA would use the NI category to exclude the words that do not exhibit clear topic patterns, so the resulting topics are more consistent and robust to the initializations of multiple MCMC runs.

We also measure the consistency of vsLDA by looking at the NI word sets over multiple runs and computing the Jaccard’s coefficient, which measures the degree of overlap of two sets by dividing the intersection by the union. Although we do not know the ‘ground truth’ of the NI word set, we certainly do not expect it to change for each run. The Jaccard’s coefficients for multiple MCMC runs are, on average, 0.83, 0.96, and 0.93 for 20NG, NIPS, SigGraph, respectively, and these values represent high consistencies over multiple runs.

## 6 Discussion

We developed a variable selection model for LDA which selects a subset of the vocabulary to better model the topics. We were motivated by the curiosity about the usual practice of using the entire vocabulary to model the topics and the ad-hoc nature of the preprocessing steps to reduce the vocabulary size. Our model, vsLDA, explicitly selects the non-informative words to exclude from the vocabulary, simultaneously with the inference of the latent topics. By only using the words that help, not hinder, the process of inferring the topics, our model combines the advantages of LDA with symmetric priors and LDA with asymmetric priors. One future direction for vsLDA is to apply it to online learning [10, 24]. Typically, in an online learning situation, the vocabulary size gets larger as more data become available, but we cannot use the entire vocabulary because it monotonically increases [14]. By using vsLDA we can control the effective size of the vocabulary. Also, vsLDA can be used for object recognition, image segmentation [23, 25], or collaborative filtering [11, 15] because vsLDA finds topics with more discriminative power. With vsLDA, we showed one way

of incorporating variable selection into LDA and improving the results, so the natural next step would be to incorporate variable selection into other topic models [1, 13, 18, 8] for improved results.

## References

- [1] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, 2004.
- [2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.
- [3] J Boyd-Graber, J Chang, and S Gerrish. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, 2009.
- [4] C. Chemudugunta and P.S.M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, page 241. The MIT Press, 2007.
- [5] Edward I. George and Robert E. McCulloch. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.
- [6] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, 2003.
- [7] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems*, 2005.
- [8] Z. Guo, S. Zhu, Z.M. Zhang, Y. Chi, and Y. Gong. A topic model for linked documents and update rules for its estimation. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.
- [9] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(1):1157–1182, 2003.
- [10] Matthew D Hoffman, David M Blei, and Francis Bach. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 23, pages 1–9, 2010.
- [11] Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22(1):89–115, January 2004.

- [12] Harold W. Kuhn. The hungarian method for the assignment problem. In *50 Years of Integer Programming 1958-2008*, pages 29–47. Springer Berlin Heidelberg, 2010.
- [13] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 577–584, New York, NY, USA, 2006. ACM.
- [14] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
- [15] Benjamin Marlin. Modeling user rating profiles for collaborative filtering. In *Advances in Neural Information Processing Systems*, 2003.
- [16] Marina Meila. Comparing Clusterings by the Variation of Information. In *In Proceedings of the 16th Annual Conference on Computational Learning Theory*, pages 173–187. 2003.
- [17] Thomas P. Minka. Estimating a Dirichlet distribution. *Technical report, Microsoft Research*, 2003.
- [18] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494. AUAI Press, 2004.
- [19] Yee Whye Teh, Matthew J. Beal, Michael I. Jordan, and David M. Blei. Hierarchical dirichlet processes. *Journal of The American Statistical Association*, 101:1566–1581, 2006.
- [20] H. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why Priors Matter. In *Advances in Neural Information Processing Systems*, 2009.
- [21] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. *Proceedings of the 26th International Conference on Machine Learning*, 2009.
- [22] C. Wang and D. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in Neural Information Processing Systems*, 2010.
- [23] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2007.
- [24] Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. *Proceedings of the 15th International Conference on Knowledge Discovery and Data mining*, pages 937–946, 2009.

- [25] Bin Zhao, Li Fei-Fei, and Eric P. Xing. Image segmentation with topic random field. In *Proceedings of the 11th European Conference on Computer vision*, pages 785–798, Berlin, Heidelberg, 2010. Springer-Verlag.